

MathDeck Overview

MathDeck is a user-friendly search engine for formulas and text in PDF. The system supports formula extraction, search, editing with LaTeX, handwriting and visual operations, and formula annotation and export.

Search Interface Features

- **Queries** combine formulas + text
- **Highlights** matched formulas + text in PDF pages
- **Import** formulas as 'chips' for search, editing, reuse
- **Cards** add titles and descriptions for formula 'chips'
- **Decks** collect searchable cards + autocompletion
- **Formula editor** with LaTeX + visual operations [1]

(1961) A fourth level of linguistic analysis

Top of a Document Hit Summary. Includes BibTeX button, ACL link & view formulas button (fx)

Collection: ACL Anthology PDFs

- 11,021 PDFs from 1952-2005 with BibTeX data
- 463,050 extracted formulas
- **Formula Extraction:** ScanSSD-XYc [2] and SymbolScraper [3] extract symbols and formula locations; QDGGGA [4] parses images into LaTeX
- **Text:** SymbolScraper provides word locations

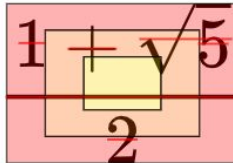


If tf term $w_{d,t}=1+\ln(f_{d,t})$ how apply idf term $w_{q,t}=\ln(\frac{N}{n_t}+1)$

HIDE RESULTS add formula... x | +

Example Search Result. Formula hits are shown in yellow, text hits in blue, and additional formulas in gray.

Example Card Deck Showing TF and IDF



Example PHOC: 3 Nested Rectangles ('R3' Level 3). Red lines: symbol 'locations'

PDF Retrieval: Formula PHOC + BM25

PDFs ranked by best-scoring page. Ranking combines **Pyramidal Histogram of Character (PHOC)** formula retrieval with BM25 for text, in *OpenSearch*

Retrieval Model Summary

- **Formula PHOC:** spatially-augmented bag of words [5]: binary vectors for symbols in overlapping regions. **R10** with concentric rectangles used (55 regions/bits)
- **Formula similarity:** cosine similarity over PHOC vectors (using fast bitwise operations)

$$\cos(\mathbf{a}, \mathbf{b})^{rank} \approx b\cos(\mathbf{a}, \mathbf{b}) = \frac{1}{\sqrt{|b|_1}} |\mathbf{a} \wedge \mathbf{b}|_1$$

- **Documents Ranked by Highest Scoring Page.** Page score is **BM25** text score + **3 times** the sum of the top match for each query formula on a page

ARQMath-2 [6] Formula Retrieval Task

R10 PHOC model competitive despite being a sparse 'spatial' bag-of-words model without unification of symbols

System	NDGG'	MAP'	P@10
R10-PHOC	0.5136	0.3005	0.4810
Approach0	0.6520	0.4710	0.6120
TangentCFT	0.6300	0.4830	0.6620

[1] Gavin Nishizawa, Jennifer Liu, Yancarlos Diaz, Abishai Dmello, Wei Zhong, and R. Zanibbi. 2020. *MathSeer: A math-aware search interface with intuitive formula editing, reuse, and lookup*. CHI 2020.

[2] Abhisek Dey and Richard Zanibbi. 2021. *ScanSSD-XYc: Faster detection for math formulas*. ICDAR 2021.

[3] Ayush Kumar Shah, Abhisek Dey, and Richard Zanibbi. 2021. *A math formula extraction and evaluation framework for PDF documents*. ICDAR 2021.

[4] Mahshad Mahdavi and Richard Zanibbi. 2020. *Visual parsing with query-driven global graph attention (QD GGA): Preliminary results for handwritten math formula recognition*. CVPR Work. 2020.

[5] Matt Langsenkamp, Behrooz Mansouri, and Richard Zanibbi. *Expanding spatial regions and incorporating IDF for PHOC-based math formula retrieval at ARQMath-3*. CLEF 2022.

[6] Behrooz Mansouri, Vit Novotny, Anurag Agarwal, Douglas W. Oard, and Richard Zanibbi. 2022. *Overview of ARQMath-3 (2022): Third CLEF lab on Answer Retrieval for Questions on Math*. CLEF 2022.